Research Professorship Proposal - Knowledge Discovery for Large Data Sets

Alina Lazăr
Department of Computer Science and Information Systems
Youngstown State University
alazar@cis.ysu.edu

Abstract

The proposed research will investigate and develop knowledge discovery tools for extracting meaningful socio-economical knowledge from data sets derived from the Current Populations Survey conducted by the U.S. Census. The survey, conducted for the past 50 years, centralizes social, demographic and economical data intended to provide policymakers and legislators with statistics for the planning and evaluation of government programs. However, the extraction of useful knowledge from such large data sets is a very demanding task that requires the use of sophisticated techniques. The data sets need to be filtered and preprocessed to eliminate irrelevant attributes and incompleteness, before building classifiers and predictors to efficiently extract information. Support vector machines have proved to be valuable in terms of prediction accuracy, but the actual knowledge is hidden in a "black box", which makes difficult to track the rules that govern the output. The present research intends to combine these methods with rough logic techniques, which while not being powerful predictors, are able to provide a set of rules which improves the user's understanding. The goal is to analyze the performance of these methods in terms of training time, correct classification rates and the knowledge extracted in a linguistic representation.

This project can generate results in less than a year providing the basic algorithms to be used in future knowledge discovery out of large real data sets. It will also provide excellent small projects opportunities to encourage undergraduate students to pursue research.

III. Body of Application

A. Narrative

1. Needs Statement

1.1 Background

In today's society the Information Technology (IT) is an increasingly integrant part of all economic, technological, educational and even cultural sectors. Through applications such as, e-commerce, networking, digital administration the IT revolution has become one of the most important factors in shaping the future of our social system. This advent is a consequence not only of the extraordinary pace of the technological advancement in the last century, but of the importance that our society gives to the storage and manipulation of its main currency, i.e. information.

A recent study, for the year 2002, conducted at University of California Berkley by Lyman and Varian [11], regarding the new information generated around the world, showed that out of the 5 extrabytes (2^{60} bytes) of data generated during last year, more than 90% were stored on magnetic media, mostly on harddrives, compared with only 7% on film and 0.01% on paper. Another survey done by Winter's Corporation [5] found that for the first time in history a decision-support database surpassed a typical transaction-processing databases in the competition for the largest and most heavily used database. The decision-support database grown 30% in the last 2 years reaching today 30 terabytes (2^{40} bytes) and it will continue to increase because of the broadband internet expansion and the new Radio Frequency Identification Devices (RFID).

However, the benefits of having an increasing amount of data safely and easily saved and readily available is shadowed by the increasingly difficult task of sorting and extracting the meaningful and useful information. While advances in basic hardware and software platforms capable to generate and store data at increasing speed is important to sustain the development pace of our society, it is equally critical to develop methods and programs sophisticated enough to transform effectively the increasing amount of data in valuable knowledge and to discover the hidden relationships within this data. This requirement has lead to the development in the computer science discipline of the data mining field, which combines methods and approaches from the fields of machine learning, databases, cluster analysis, statistics and visualizations. While machine learning studies have already tackled problems such as data structures and algorithms for knowledge discovery, they did not put a specific emphasis on efficiency and scalability, because the data sets used are typically small laboratory generated ones. Nevertheless, these are central issues in data mining since in this case the data sets large ones acquired from the real world. Besides their size these sets are generally plagued by missing values, dynamic effects and noise, which require different approaches from the machine learning methods.

One of the main issues in today's data mining field is the communication gap between the data mining scientist, who elaborates algorithms and methods for knowledge extraction, and the scientists in other various fields, who produce the actual data sets and have to analyze and validate the outcome of any data mining endeavor. Most of the researched algorithms and methods, while providing good classification and prediction tools, fail to explain the decisions to people with insufficient experience in data mining techniques. However, many fields, like the medical or the security one, require that a decision system should be able to explain and justify the decisions especially when it provides an unexpected solution to a new problem. One proved method to circumvent this problem is using rule induction methods such as decision trees or rough sets [8, 9, 10]. These methods allow not only the results of the data mining to be extracted, but also the rules that determined these results.

The present research will use the Current Population Survey (CPS) database provided by the U.S. Census Bureau [4, 14]. The CPS survey was conducted for more than 50 years and collects information about the social, demographic and economic characteristics of the labor force 16 years and older of the U.S. population. The data collected each month is used to compute reports about employment, unemployment, earnings. It also includes statistics about various social factors from voting to smoking. The government policymakers and legislators use the statistics generated from the CPS data as indicators about the economic and social situation and for the planning and evaluation of many government programs. The data is publicly available and free of charge, fact that encouraged its use in various social and economic studies [6, 12]. However, due to the large number of variables included and its implicit large size it is farfetched to believe that its entire value has been fully exploited. This has motivated its active use by the machine learning and knowledge discovery communities, as a platform for testing various data mining methods including, neural networks, nearest neighbor, decision tree and lately support vector machines.

Two data sets were previously extracted from the CPS data and posted on the University of California Irvine (UCI) repository [13]. Both of them were tailored to predict if a person income is more or less than 50K. The first data set, posted on the UCI machine learning repository and named "adult dataset", was extracted from the 1994 CPS data. Records with unreal values were eliminated and finally the 48,842 instances were divided into two files: a training file and a testing one. Fourteen attributes, eight categorical and six continuous were chosen. They were age, work class, weight, education, years of education, marital status, occupation, relationship, race, sex, capital gain, capital loss, hours per week and native country. The six continuous attributes were quantified into quintiles before running any algorithm. The second data set named "Census Income Database", posted on the UCI KDD repository, with 1,999,523 instances is larger than the adult dataset. It was extracted from the 1994 and 1995 CPS data and contains 41 demographic and economic related variables. These attributes include the majority but not all of the 14 attributes included in the adult data set. Another difference between the two data sets is the decision variable provided for the classification problem. For the adult dataset the decision was drawn from the "adjusted gross income" versus the "total personal income" for the census income database

1.2 Statement of the Problem

The goal of the proposed research is to update the adult and census income data sets. The U.S. Census Bureau conducts the Current Population Survey every month, and in conclusion a large quantity of data was generated since 1994-1995 when the two data sets were extracted. A similar data set with the adult dataset can be derived each month because the monthly survey covers 50K households. For a more realistic classification task, we believe that the 50K threshold should be replaced with the median income for the last year (~42K). We plan to start with a complementary set including all the variables used in the previous two studies and apply feature selection algorithms [1] to find the most important attributes for predicting income. Nevertheless, it is difficult to draw a clear line between the feature selection task and the classification procedure. Because of that an important part of the research will be to explore feature selection methods classifiers and rule induction approaches for income prediction. The exploration will be done via a systematically evaluation of the performances (training time, accuracy) for the chosen method.

Several machine learning techniques have been previously applied to classify and predict the income. These methods include neural networks, decision tree, nearest neighbor, support vector machine. However none of them is able to output rule sets which are critical for economy and social scientists to evaluate the soundness of the results. We would like to combine the best support vector machine existent with rough sets logic in order to build a good classifier which will also justified its predictions using a rule set.

The support vector machine (SVM) methods [3, 6, 7, 12] build a hyperplan or a decision surface, based on the risk minimization principle, to maximize the separation between two classes of examples. At each step to the SVM is given a sequence of hypothesis spaces of increasing complexity and the training error is computed. From the hypothesis sequence only one is chosen, the one which minimizes the training errors. From one step to the other the feature weights are modified. In order to deal with nonlinear decision, surfaces kernel functions are used.

The rough sets methods [8, 9, 10] are based on equivalence relations, are setting approximations and are able to build optimal sets of decision rules. Because of their three-valued simplicity: lower, upper and boundary approximation sets, they work well on discrete and categorical data. They also can be useful with missing data, change of scale and problems where membership grades are hard to define.

1.3 Significance and Appropriateness

In the new era called "information age", the rapid evolution of our society is based on knowledge. Information means power today. Whoever is able to derive or extract the essential knowledge from a variety of data sources wins. The knowledge usually takes the form of trends, patters or regularities in data files. This process is very expensive computationally speaking and it can not be done in real time. Another problem is that the knowledge discovery process can very easily lead to meaningfulness patterns if it is not carefully planned. Therefore, new tools that can extract and in the same time increase the understandability of the decision making process are needed.

1.4 Innovativeness and Generalizability

As we mention above building strong decision support systems is not an easy task. We plan to bring novelty in this process by combining together good classification and prediction tools with rule based instruments capable to generate some useful explanations for the extracted knowledge.

Beside the current population survey other data from censuses and surveys containing important socio-economic information is available at the U.S. Census Bureau (i.e. survey of income and participation, mortality, national health and nutrition examination survey). All the steps and algorithms applied to the current population survey can be extended to any other data. Also, successful machine learning and data mining algorithms can be transferred to other applications. For example, after the recent development of microarray technology a lot of the research effort of the community were directed towards predicting and diagnosing diseases, such as cancer..

1.7 Relationship to Applicant's Current Research

My interest in knowledge discovery and data mining dates back to my graduate studies. One of my first papers [10] in this field won the Best Paper Award at the ANNIE'99 conference for theoretical developments in computer intelligence. The aim of the paper was to describe a new method for making the knowledge embedded in a trained neural network comprehensible and thus transform neural networks form "black boxes" into a powerful acquisition tool.

Some of my other papers [9] and my doctoral thesis covered the extraction of meaningful knowledge from an archaeological data set and used it into a multi-agent simulation for the formation of the archaic state in the Valley of Oaxaca, Mexico. Our research goal was to integrate evolutionary learning tools into the data mining algorithms and apply them to the large-scale spatial-temporal archaeological data set produced by surveys. After scanning the data we followed the steps of the knowledge discovery process. Most of the results fit in with the empirical findings of the anthropologists. We would like to use the same knowledge discovery process in the case of the CPS data, which has common trends with the archaeological data already studied.

1.8 Relationship to Departmental/College Missions and Goals

As stated in the vision statement of the Computer Science and Information, applied research and interdisciplinary projects are highly encouraged: "We will become a department noted for its commitment to applied research …"; "The Department is an important academic resource for the campus and the community"; "The Department supports interdisciplinary programs and forms symbiotic relationships with other disciplines when appropriate and possible".

Currently, the Department of Computer Science and Information Technology is introducing a Master Program in its curriculum. It is envisioned that part of the work dedicated to this project will contribute to the development of a graduate level classes like "Data Mining" and "Advanced Databases". In the same time the research subject involved might be very attractive for prospective graduate students. It is also my intention to involve undergraduate students especially in processing the data and running algorithms.

2. Statement of Objectives

The main objectives for the proposed research are:

1. To develop useful automatic procedures for building data sets out of the CPS survey.
2. To integrate powerful data mining tools with different capabilities such as support vector machine and rough sets.
3. To extract “knowledge for use” that can increase the understandability of important decision makers.
4. To disseminate the data sets and the results of the knowledge discover in the Data Mining and Machine Learning communities.

3. Procedures

One of the most important problems in data analysis relates to the dimensionality of the data. Because many data analysis techniques involve exhaustive search over the object space, they are very sensitive to the size of the data in terms of time complexity and it is hard to generate compact rules. The solution is to reduce the search space horizontally (in terms of records or objects) and vertically (in terms of fields or attributes or variables), and to use heuristics to guide the search through the large space of possible combinations of attributes values and classes.

Beside their size real data sets tend to contain an uncertainty dimension. Errors which can occur during data collection or data entry are referred as noise in the data. It is also possible that the data set can have missing attribute values. In this case, the objects containing missing attributes values can be discarded or the missing values can be replaced with the most common values. Another problem is that the available knowledge in many situations is incomplete and imprecise. This means that sometimes the attribute values for a set of objects are not sufficient and precise enough to differentiate between classes of objects. Many different ways of representing and reasoning about uncertainty have been developed in the artificial intelligence field. These theories includes: belief networks, non monotonic logic, fuzzy sets along with fuzzy logic and rough sets. The rough sets approach, which will be used in the present research, provides a lower and upper approximation in terms of sets, belonging to a concept depending on how the relationship between two partitions of a finite universe is defined.

In addition to incomplete data, a data set, especially one collected through large surveys, may contain redundant or insignificant attributes with respect to the problem, and variables that are obscure, and have non-interesting relationships. This case might arise in several situations. For example, redundant attributes may result from combining relational data tables. Solution to this problem exists, in the form of feature selection algorithms [1,8], such as the reduct computation in the rough sets case. After discharging the irrelevant attributes we may have horizontal or object related redundancy. This can be solved by applying horizontal pruning methods, or merging identical objects.

A fundamental characteristic of real world data sets is that they are dynamic, which means that their contents are often changing over time. There are two important aspects of these problems. First, the run time efficiency of the knowledge model becomes very important. Second, the knowledge model will no longer be static, but should have the capability of evolving as data changes over time.

The main steps to be followed in elaborating the knowledge discovery process based on the previous observations can be summarized as follows:

- *Sampling and selection* - the irrelevant attributes are removed and the selected data is represented as a two-dimensional table.

- *Preprocessing* - if the selected table contains missing values or empty cell entries, the table must be preprocessed in order to remove some of the incompleteness. Statistics should be run to obtain more information about the data.

- *Transformation* - for example, measurement attributes should be discretized, and used instead of exact observations. Categorical data may be recoded to provide a consistent interpretation. Also, the decision variables should be identified. After this step the data becomes more qualitative than quantitative.

- *Training and validation sample* - the initial table is divided into at least two subtables. One will be used in the training step, the other in the validation or testing step.

- *Develop the mode* - knowledge discovery techniques are applied to the training data in order to generate a set of hypothesized relations. Following the rough sets methodology, the full set of reducts is computed, a set of minimal reducts is chosen, and the data table is vertically pruned. Then the object related reducts are computed and the exhaustive decision rule system is generated. At the end a pruning method for the decision rule set is applied in order to obtain a good decision system, with a good balance between the number of rules and the accuracy of the classifications.

- *Interpretation and evaluation* - the validation or test data set is then used to test the classificatory performance of the new model. Also, if it is a rule-based model, it can be checked by specialists, in order to understand the data sets, understand and explain dependencies between values of attributes and definitions of decision classes. The expert will check the decision rule system.

3. 2. Project Plan and Timeline

The main part of the proposed project will be developed in the academic year 2004 – 2005. The time line and the main plan of the project are summarized in Table 1. However, due to the richness and continuous evolution of the data contained in the Current Population Survey (CPS), it is expected that the present proposal will establish the base for other research projects beyond the period covered by the following timeline. The plan is to submit a proposal, having as focus this research, for the NSF Grant "Information and Data Management", which is due January 8, 2004.

4. Statement of Evaluation

The timeline illustrated in Table 1 sets the template for the progress assessment of the proposed research. The critical evaluation stages will include:

- efficiency and accuracy of the new algorithms when compared with old packages, in terms of speed and validity of results. The embedded chronological nature of the Current Population Survey data, offers a valuable assessment opportunity by allowing results obtained from older data sets to be compared to real data obtained at a later time.
- analysis of the required time for meaningful knowledge extraction. Handling of large data sets may require transition to high performance computing platforms, i.e. parallel computers.
- response of the interested governmental, social and economic institutions to the disseminated results and rules extracted.

Task Name	**2004**				**2005**							
	S	O	N	D	J	F	M	A	M	J	J	A
Literature review research about the Current Population Survey data, support vector machine techniques and rough sets algorithms.	◀	▬	▬									
Developing programs that automatically test new data mining algorithms against developed packages for classification and prediction	◀	▬	▬	▬	▬							
Full scale integration off the successful algorithms to large data sets. Efficiency evaluation on PC and parallel computing platforms.			▬	▬	▬	▬	▬	▬	▬	▬	▬	▶
Dissemination of results through papers and communications at specific conferences.				▬	▬	▬	▬	▬	▬	▬	▬	▶
Evaluation of the applicability of the developed algorithms to other attributes contained in the Current Population Survey data.									▬	▬	▬	▶

Table 1. Timeline of the Knowledge Discovery for Large Data Sets proposal.

5. Plans for Dissemination of Results

The findings of this research will be made available through papers and reports in dedicated journals of data mining, machine learning and artificial intelligence such as: *International Journal of Intelligent Systems* and *IEEE Transactions on Knowledge and Data Engineering.* Also, the participation at conferences and workshop will constitute a good opportunity to directly make available the obtained knowledge to the interested audiences. The first new results will be communicated in the *QUEST 2004* and *2005 Workshops* hosted by Youngstown State University. National and international conferences in which these results are expected to stir interest are: the *KDD-05 International Conference on Knowledge Discovery and Data Mining*, the *2005 SIAM International Conference on Data Mining* and the *2005 National Conference on Artificial Intelligence*. The new large data sets will be submitted to the UC Irvine Knowledge Discovery in Databases (KDD) Archive, a new online repository of large data sets. The mission of this repository is to provide researchers within the data mining field with large and complex data sets.

B. Availability of needed resources

In the first phase of the project the selection and preprocessing of datasets from the Current Population Survey (CPS) database will be performed on the existent computing platforms at the CSIS department at YSU. In the second phase the results of our algorithms will be compared against results obtained using free publicly available software platforms for classification and prediction algorithms. As the project moves in its third phase, dedicated to the full implementation of the successful algorithms identified in the second phase, the real possibility of the datasets being to large for efficient analysis on PC platforms, has to be considered. If such a problem is identified a transition to high performance, i.e. parallel computing, platforms is envisioned. A supercomputer cluster Colony existent at YSU provides researchers from different fields with a unique opportunity for highly computationally demanding scientific applications. The supercomputer, built in 2002, is a Linux cluster with a total of 32 processors (16 servers with two processors each) that YSU earned from the Ohio Supercomputer Center. Also the Ohio Supercomputer Centre provides computing time through grants to any researcher in the state of Ohio. This facility has been already used in the context of a different research proposal dedicated to agent-based simulations, and the gained experience is expected to provide a smooth transition for this specific project.

The main software platform that will be used in the proposed research is C++ available through a university license. For the rough sets the Rosetta free software will be used. Classification support vector machine implementations, such as, LIBSVM, SVMlight and SVMFu3 are readily available as freeware. For two other useful software packages, Matlab and the statistical package SPSS, the best academic deal will be negotiated.

IV Appendix A - References

[1] Bradley P. S. and Mangasarian, O. L., "Feature Selection via Concave Minimization and Support Vector Machines". Mathematical Programming Technical Report 98-03, February 1998. "Machine Learning Proceedings of the Fifteenth International Conference(ICML '98)", J. Shavlik, editor, Morgan Kaufmann, San Francisco, California, 82-90, 1998.

[2] Cho, S.B., Won, H.H., "Machine Learning in DNA Microarray Analysis for Cancer Classification". First Asia-Pacific Bioinformatics Conference (APBC 2003), February 4-7, 2003, Adelaide, Australia, Australian Computer Society, CRPIT, Vol. 19, pp:189-198

[3] Cristianini N. and Shawe-Taylor, J*., An Introduction to Support Vector Machines*, Cambridge University Press; 2000

[4] Current Population Survey, U.S. Census Bureau. Retrieved from http://www.bls.census.gov/cps/cpsmain.htm on 11/16/03.

[5] Hicks, Matt **"**Survey: Biggest Databases Approach 30 Terabytes." eWeek, November, 2003. Retrieve from http://www.eweek.com/ on 11/16/2003

[6] Joachims, T., "Making large-scale SVM learning practical, " *in Advances in Kernel Methods - Support Vector Learning*, (B. Schölkopf, C. Burges, and A. Smola, eds.), pp. 169-184, MIT Press, Cambridge, MA, (1999).

[7] Komarek, P. and Moore, A. "Fast Logistic Regression for Data Mining, Text Classification and Link Detection" submitted to NIPS, June 2003

[8] Komorowski, J. and Øhrn, A., *"*Modelling Prognostic Power of Cardiac Tests Using Rough Sets", invited paper, Special Issue on Intelligent Prognostic Methods in Medicine, *J. of Artificial Intelligence in Medicine* (1999) Vol. 15:2, pp. 167-191, Elsevier Science.

[9] Lazar, A. and Reynolds, R.G., "Evolution-based Learning of Ontological Knowledge for a Large-scale Multi-agent Simulation", at *The Fourth International Workshop on Frontiers in Evolutionary Algorithms* (FEA 2002), Research Triangle Park, North Carolina, USA, March 8-13, 2002

[10] Lazar, A. and Sethi, I.K., "Decision Rule Extraction from Trained Neural Networks Using Rough Sets" in *Intelligent Engineering Systems Through Artificial Neural Netowks* (Dagli, C.H., Buckzak, A.L., and Ghosh, J., eds.) vol. 9, (New York, NY), pp. 493-498, ASME Press, Nov. 1999

[11] Lyman, Peter and Varian, Hal R., "How Much Information", 2003. Retrieved from http://www.sims.berkeley.edu/how-much-info on 11/16/2003.

[12] Platt*,* J.C., "Fast Training of Support Vector Machines using Sequential Minimal Optimization" *in Advances in Kernel Methods - Support Vector Learning*, (B. Schölkopf, C. Burges, and A. Smola, eds.), pp. 185-208, MIT Press, Cambridge, MA, (1999).

[13] The UCI ML and KDD Archive. Retrieve from http://www.ics.uci.edu/~mlearn/MLRepository.html and http://kdd.ics.uci.edu on 11/16/03, Irvine, CA: University of California, Department of Information and Computer Science.

[14] U.S. Census Bureau, United States Department of Commerce. Retrieved from http://www.census.gov/ on 11/16/03.